TITLE OF THE INVENTION

Method and apparatus for multiple byte or page mode programming and reading and for erasing of a flash memory array

INVENTORS

Kyung Joon Han of Palo Alto, California

Dung Tran of San Jose, California

Steven W. Longcor of Mountain View, California

Steve K. Hsia of San Jose, California



CROSS-REFERENCE TO RELATED APPLICATION

**[0001]** This application claims the benefit of U.S. Provisional Patent Application No. 60/291,156, filed May 14, 2001 (Hsia et al., "Apparatus and method for multiple byte or page mode programming or erasure in a nonvolatile flash memory array," Attorney Docket No. 11030.00), which is hereby incorporated herein by reference in its entirety.

BACKGROUND OF THE INVENTION

**[0002]** Field of the Invention

**[0003]** The present invention relates to semiconductor memory, and more particularly to nonvolatile semiconductor memory that is programmable as well as erasable.

**[0004]** Description of the Related Art

**[0005]** Nonvolatile memory retains stored data when power is removed, which is required or at least highly desirable in many different types of computers and other electronic devices. Some types of nonvolatile memory are capable of being repeatedly programmed and erased, including erasable programmable read only semiconductor memory generally known as EPROM, and electrically erasable programmable read only

semiconductor memory generally known as EEPROM. EPROM memory is erased by application of ultraviolet light and programmed by application of various voltages, while EEPROM memory is both erased and programmed by application of various voltages. EPROMs and EEPROMs have suitable structures, generally known as floating gates, that are charged or discharged in accordance with data to be stored thereon. The charge on the floating gate establishes the threshold voltage, or $V_T$, of the device, which is sensed when the memory is read to determine the data stored therein.

**[0006]** An illustrative well known type of compact floating gate EEPROM cell structure is the stacked gate structure shown in FIG. 1. A floating gate 14, typically a doped polysilicon layer, is sandwiched between two insulator layers 12 and 16, typically oxide. The top layer of the stack is a control gate electrode 10, typically a doped polysilicon layer. The stacked gate structure is shown symmetrically overlying part of a heavily doped *n*+ source region 20 and a heavily doped *n*+ drain region 22, as well as a channel region between the source region 20 and the drain region 22. The channel region is part of a *p*-well 28, which also contains the source region 20, the drain region 22, and a heavily *p*+ doped contact region 24. The *p*-well 28 typically is contained within an *n*-type substrate or within an *n*-well such as shown at 30, which also contains a heavily *n*+ doped contact region 26. The *n*-well 30 is in turn contained in the *p*-type substrate 32. Many variations in the floating gate EEPROM cell structure are known, and include asymmetrical stacked gate structures, split gate structures, and so forth. Moreover, although the structure of FIG. 1 is an *n*-channel enhancement mode device, nonvolatile memory cells may be fabricated as either *n*-channel or *p*-channel devices or as enhancement or depletion mode devices.

**[0007]** As is typical of nonvolatile memory cells that are capable of being repeatedly programmed and erased, the various functions of the EEPROM stacked gate memory cell of FIG. 1 are controlled by applying various bias voltages. The voltage applied to the control gate is $V_G$, the voltage applied to the source is $V_S$, the voltage applied to the drain is $V_D$, the voltage applied to the *p*-well 28 is $V_P$, the voltage applied to the *n*-well 30 is $V_N$, and the voltage applied to the *p*-type substrate 32 is $V_B$ (not shown). Typically the substrate 32 is grounded, *i.e.* $V_B = 0V$. Typically writing or

programming the memory cell means adding negative charge to the floating gate while erasing the memory cell means removing negative charge from the floating gate, but the charged state can be considered the erased state if desired. Other voltages are applied to read the charge state of the memory cell by detecting the threshold voltage $V_T$ of the memory cell, which ideally is done without disturbing the charge state.

**[0008]** Depending to some extent on device characteristics, the stacked gate transistor of FIG. 1 may be programmed by moving electrons to the floating gate 16 using Fowler-Nordheim ("FN") tunneling or electron injection. Electron injection typically is done using channel hot electron injection ("CHE") or channel-initiated secondary electron injection ("CISEI").

**[0009]** The EEPROM stacked gate memory cell of FIG. 1 may be used in a variety of memory array architectures, including common ground arrays as well as virtual ground arrays. A memory is formed by combining a memory array with well known circuitry such as control logic, address decoders, sense amplifiers, and power supplies. An example of a memory 40 having an flash memory array 54 of such individual cells is shown in FIG. 2. Various read, erase and program voltages are furnished by suitable power supplies (not shown). A serial memory address ADDR is latched into an address latch 44, decoded for its row and column information (X and Y) by X decoder 48 and Y decoder 46, and applied to the memory array 54 to access the selected row and column. If the operation is a program operation, the data to be written is temporarily stored in I/O buffer 50 as it is written to the memory array 54. If the operation is a read, the selected bits are sensed by sense amplifier 52 and then temporarily stored in the I/O buffer 50, where they are accessible to external circuits.

**[0010]** For many memory applications, one desires to read and program multiple bytes of the memory array 54 simultaneously, or even an entire page of the memory array 54. Similarly, one may desire to erase multiple bytes or even an entire page of the memory array 54 at one time, or even multiple pages or the entire memory. To facilitate erasing, programming and reading multiple bytes or even an entire page, each row of the memory array or perhaps adjacent rows may correspond to a page of memory. A sector of

memory may contain several pages. Such memory is known as "flash" memory because of the large number of bits that can be erased or programmed simultaneously.

**[0011]** One type of conventional flash memory uses FN tunneling for both erasure and programming. Unfortunately, programming using FN tunneling from the drain edge to the floating gate is relatively slow. Transistors using FN programming generally requires a longer channel length, leading to larger cell size. FN programmed memories also require bit-latch circuitry, which increases the size of the memory chip.

**[0012]** Another type of conventional flash memory uses CHE for programming. CHE programming is fast relative to FN programming. Unfortunately, the high drain voltage and programming current required by CHE renders the technique disadvantageous for use in low power applications, and severely limits the number of bits that can be programmed at one time. Simultaneous multiple byte programming is difficult to perform, as a practical matter.

**[0013]** While multiple byte programming and page mode programming of a CHE type memory can be achieved by repeated programming groups of bits until the desired amount of memory is programmed, the approach can result in an unfavorable condition known as program-disturb. Program-disturb is related to the voltage conditions that occur in the part of the memory that is not being programmed while another part of the memory is being programmed. These voltage conditions cause multiple minute shifts in the threshold voltage of the memory cells that are not being programmed, which occur as other parts of the memory are being programmed. A similar problem occurs during read-out of data. Read voltages applied to the nonvolatile cells, including both the addressed cells and some of the cells that are not addressed, can induce a threshold voltage shift in these cells. While program-disturb and read-disturb can be avoided by the use of an isolating select transistor in each memory cell, such transistors are undesirable insofar as they cause an increase in the size of the memory cell and a corresponding decrease in the memory array density.

**[0014]** A technique is known that uses negative substrate biasing of the flash memory cells to overcome some of the disadvantages of conventional CHE. An example of this technique is disclosed in United States Patent No. 5,659,504, which issued August

19, 1997, to Bude et al. and is entitled "Method and Apparatus for Hot Carrier Injection." The Bude et al. programming technique, which is referred to as channel-initiated secondary electron injection ("CISEI"), uses a positive bias voltage of about 1.1 volts to about 3.3 volts at the drain and a negative bias voltage of about -0.5 volts or more negative at the substrate, with the source at zero volts. The source-drain voltage causes some channel hot electron generation while the substrate bias promotes the generation of a sufficient amount of secondary hot electrons having a sufficient amount of energy to overcome the energy barrier between the substrate and the floating gate. The secondary hot electrons are primarily involved in charging the floating gate. The programming of the flash memory array using CISEI transistors is relatively quickly achieved with low programming current, low drain voltage, and smaller cell size (shorter channel length) relative to flash memory arrays using CHE transistors. However, simultaneous multiple byte programming and page mode programming are still difficult to achieve. Unfortunately, as in the case the CHE memory array, the use of isolating select transistors in CISEI memory cells increases their size, and the technique of repeated programming groups of bits until the desired amount of memory is programmed can cause program-disturb.

**[0015]** While CHE and CISEI cell programming is faster that FN cell programming, multiple byte programming and page mode programming of CHE and CISEI memory arrays remains problematical. FN tunneling remains a popular choice in flash memory for erase operations.

## BRIEF SUMMARY OF THE INVENTION

**[0016]** We have found that flash memory suffers disturbance of the floating gate potential especially during page mode programming operations, and may also suffer disturbance of the floating gate potential during read operations. We have also found that the relatively thin high quality tunnel oxide commonly found in EEPROM memory cells has a shortened lifetime because of the high fields that occur across the tunnel oxide during the FN erase operations.

**[0017]** These and other disadvantages are overcome individually or collectively in various embodiments of the present invention. For example, one embodiment of the present invention is a method of programming a memory array that comprises a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate overlying a channel and being programmable using channel hot electron injection. The method comprises applying a first voltage to the channels; establishing a voltage differential across the respective channels of at least a first and a second of the memory cells, the potential differential being sufficient to generate channel hot electrons in the respective channels thereof; applying a second voltage to the gate of the first memory cell, the second voltage having a polarity and magnitude relative to the first voltage sufficient to attract the hot electrons and change the threshold voltage of the first memory cell to a programmed state; and applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to repel the hot electrons and deter change in the threshold voltage of the second memory cell.

**[0018]** Another embodiment of the present invention is a method of reading a memory array that comprises a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate overlying a channel and being programmable using channel hot electron injection. The method comprises applying a first voltage to the channels; establishing a voltage differential across the respective channels of at least a first and a second of the memory cells; applying a second voltage to the gate of the first memory cell, the second voltage in conjunction with the voltage differential causing a reading of the first memory cell; and applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to repel electrons generated in the channel of the second memory cell due to the voltage differential and deter change in the threshold voltage of the second memory cell.

**[0019]** Yet a further embodiment of the present invention is a method of erasing a memory array that comprises a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate

overlying a tunnel oxide and a channel. The method comprises applying a first voltage to the channels; applying a second voltage to the gate of the first memory cell, the second voltage having a polarity and magnitude relative to the first voltage sufficient to drive electrons from the floating gate through the tunnel oxide into the channel and change the threshold voltage of the first memory cell to an erased state; and applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to reduce a magnitude of an electric field through the tunnel oxide arising from stored charge on the floating gate.

**[0020]** Another embodiment of the present invention is a memory comprising a memory array having a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate overlying a channel and being programmable using channel hot electron injection; a voltage source for applying a first voltage to the channels; a voltage source for establishing a voltage differential across the respective channels of at least a first and a second of the memory cells, the potential differential being sufficient to generate channel hot electrons in the respective channels thereof; a voltage source for applying a second voltage to the gate of the first memory cell, the second voltage having a polarity and magnitude relative to the first voltage sufficient to attract the hot electrons and change the threshold voltage of the first memory cell to a programmed state; and a voltage source for applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to repel the hot electrons and deter change in the threshold voltage of the second memory cell.

**[0021]** Yet another embodiment of the present invention is a memory comprising a memory array having a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate overlying a channel and being programmable using channel hot electron injection; a voltage source for applying a first voltage to the channels; a voltage source for establishing a voltage differential across the respective channels of at least a first and a second of the memory cells; a voltage source for applying a second voltage to the gate of the first memory cell, the second voltage in conjunction with the voltage differential

causing a reading of the first memory cell; and a voltage source for applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to repel electrons generated in the channel of the second memory cell due to the voltage differential and deter change in the threshold voltage of the second memory cell.

**[0022]** A further embodiment of the present invention is a memory comprising a memory array having a plurality of memory cells coupled to a plurality of word select lines, each of the memory cells having an adjustable threshold voltage and a gate overlying a tunnel oxide and a channel; a voltage source for applying a first voltage to the channels; a voltage source for applying a second voltage to the gate of the first memory cell, the second voltage having a polarity and magnitude relative to the first voltage sufficient to drive electrons from the floating gate through the tunnel oxide into the channel and change the threshold voltage of the first memory cell to an erased state; and a voltage source for applying a third voltage to the gate of the second memory cell, the third voltage having a polarity and magnitude relative to the first voltage sufficient to reduce a magnitude of an electric field through the tunnel oxide arising from stored charge on the floating gate.

## BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

**[0023]** FIG. 1 is a cross-section drawing of a stacked gate EEPROM cell of the prior art.

**[0024]** FIG. 2 is a schematic block diagram of a memory device of the prior art.

**[0025]** FIG. 3 is a schematic circuit diagram of common ground NOR array of memory cells.

**[0026]** FIG. 4 is a schematic circuit diagram of a virtual ground array of memory cells.

**[0027]** FIGS. 5 – 10 are cross-section drawings of an EEPROM cell suitable for CHE programming, as exposed to various CHE programming bias voltages.

**[0028]** FIGS. 11 – 13 are cross-section drawings of an EEPROM cell suitable for CHE programming, as exposed to various Fowler-Nordheim erasure bias voltages.

**[0029]** FIG. 14 is a table showing various CHE bias voltages suitable for application to the memory cells of the memory arrays of FIGS. 3 and 4.

**[0030]** FIG. 15 is a graph showing programming time for a halo-implanted stacked gate transistor using CHE programming.

**[0031]** FIG. 16 is a graph showing programming current for a halo-implanted stacked gate transistor using CHE programming.

**[0032]** FIG. 17 is a graph showing erase time for a halo-implanted stacked gate transistor using FN tunneling for erasure.

**[0033]** FIG. 18 is a graph comparing various erase disturb characteristics for a halo-implanted stacked gate transistor using FN tunneling for erasure.

## DETAILED DESCRIPTION OF THE INVENTION, INCLUDING THE PREFERRED EMBODIMENT

**[0034]** EEPROM cells of the stacked gate type shown in FIG. 1 are used in a variety of different types of flash array architectures where they are subject to disturbance of the floating gate potential during CHE page mode programming operations, and may also suffer disturbance of the floating gate potential during read operations.

**[0035]** A common ground NOR array 100 is shown in FIG. 3. Simplified for clarity, the array 100 illustratively is arranged in M rows and N columns of memory cells, and each individual memory cell 111-114, 121-124, 131-134 and 141-144 is a stacked gate type of cell such as that shown in FIG. 1 but preferably having a halo implant in a manner well known in the art. While a variety of other type of nonvolatile memory cells using floating gates and other classes of nonvolatile memory cells using charge trapping like the MNOS (Metal-Nitride-Oxide-Semiconductor) device to set the threshold voltage $V_T$ of the cell may also be used, the use of the stacked gate type cell permits the memory array 100 to have a high integration density. Indeed, the integration density of flash

EEPROM memory using the stacked gate structure is quite large, even in a NOR memory architecture. For example, memory fabricated using 0.18 μm processes can have storage capacity as large as 128 megabits, with a single row having as many as 5000 memory cells.

**[0036]** In the illustrative memory array 100, the memory cells reside at respective row-column cross points and large groups of the memory cells share common source lines. The row lines $R_1$, $R_2$, ..., $R_{(M-1)}$ and $R_M$ are the word lines of the memory array 100, and the column lines $C_1$, $C_2$, ..., $C_{(N-1)}$ and $C_N$ are the bit lines of the memory array 100. Word line $R_1$ is connected to the control gates of transistors 111, 112, 113 and 114. Similarly, word line $R_2$ is connected to the control gates of transistors 121, 122, 123 and 124; word line $R_{(M-1)}$ is connected to the control gates of transistors 131, 132, 133 and 134; and word line $R_M$ is connected to the control gates of transistors 141, 142, 143 and 144. Column line $C_1$ is connected to the drains of transistors 111, 121, 131 and 141. Similarly, column line $C_2$ is connected to the drains of transistors 112, 122, 132 and 142; column line $C_{(N-1)}$ is connected to the drains of transistors 113, 123, 133 and 143; and column line $C_{(N)}$ is connected to the drains of transistors 114, 124, 134 and 144. Additional operating voltage is brought to groups of the memory cells on the source lines $S_1$ and $S_{(M/2)}$, which are connected to the sources of, respectively, transistors 111-114 and 121-124, and transistors 131-134 and 141-144. The source lines $S_1$ and $S_{(M/2)}$ may be commonly connected, connected in groups, or individually controllable, as desired. For purposes of clarity, a page of the memory array 100 is considered a row of memory cells, although a page may be differently defined for other memory array architectures. For purposes of clarity, the sector aspect of the architecture is not shown, although a sector illustratively has 32, 64 or more word lines. Illustratively, the memory array 100 has a storage capacity in the range of 16 to 128 Mb, with a single row having as many as 2K, 4K or even 8K memory cells. It will be appreciated that more complex arrangements such as sub-bit lines and substrate block isolation may be used as desired to enhance certain aspects of the memory, in a manner well known in the art.

**[0037]** A great many array architectures and nonvolatile semiconductor memory devices have been developed based on virtual ground contactless array architecture that

can achieve even higher memory density levels than the NOR array, as exemplified by the following patents: United States Patent No. 6,175,519, issued January 16, 2001 to Lu et al.; United States Patent No. 5,959,892, issued September 28, 1999 to Lin et al.; United States Patent No. 5,646,886, issued July 8, 1997 to Brahmbhatt; United States Patent No. 5,418,741, issued May 23, 1995 to Gill; and United States Patent No. 5,060,195, issued October 22, 1991 to Gill et al.

**[0038]** FIG. 4 shows a portion of the core of a simple virtual ground contactless array architecture 200 that uses a cross-point array configuration defined by, illustratively, buried n+ diffusions 210, 211, 212, 213 and 214 that form the bit lines and source lines $BL_{+2}$, $BL_{+1}$, $BL_0$, $BL_{-1}$, and $BL_{-2}$, and $WSi_2$/Poly control gate wordlines 220, 230 and 240, or $WL_{+1}$, $WL_0$ and $WL_{-1}$. Due to elimination of the common ground line and the drain contact in each memory cell, extremely small cell size is realized. Programming, erasing and reading of the memory cells is obtained by the use of asymmetrical floating gate transistors 221-224, 231-234, and 241-244, and suitable source and drain decoding. Various well know measures may be taken to improve performance, such as the use of metal lines to periodically connect to the bit lines to reduce bit line resistance, the use of block select transistors to control the various voltages on segmented bit lines, and so forth.

**[0039]** In practice, memory arrays such as the arrays 100 and 200 are organized into bytes and pages and redundant rows (not shown), which may be done in any desired manner. Complete memories include well known elements such as sense amplifiers, pull-up circuits, word line amplifiers, sense amplifiers, decoders, and voltage circuits, which are omitted from FIGS. 3 and 4 for clarity.

**[0040]** A variety of processes for fabricating arrays of memory cells, including halo-implanted stacked gate cells, are well known in the art. For example, one suitable process for fabricating a NOR array of stacked gate cells such as shown in FIG. 1 is the ETOX™ memory technology, which is widely described in the literature, including various publications of Intel Corporation of Santa Clara, California, and which is available as a fabrication service from various semiconductor device manufacturers. Virtual ground processes are widely described in the literature and are available as

fabrication services from various semiconductor device manufacturers, including National Semiconductor Corporation of Santa Clara, California, and Macronix International Co., Ltd. of Hsinchu, Taiwan.

**[0041]** Preferably, the memory cells in the illustrative arrays 100 and 200 are designed to be erased using Fowler-Nordheim ("FN") tunneling, and programmed using channel hot electron injection ("CHE"). Advantageously, CHE programming is significantly faster on a single cell basis than FN tunneling. Advantageously, the use of FN tunneling through the channel area for erase allows single or multiple pages to be erased with low relatively low power. An illustrative stacked gate halo-implanted transistor for the memory arrays 100 and 200 has the following exemplary basic characteristics: a grown tunnel oxide having a thickness of about 9 nm to 10 nm, an oxide-nitride-oxide ("ONO") insulator between the control gate and the floating gate having an effective thickness of about 14nm to 16nm, a phosphorus doped polysilicon floating gate having a thickness of about 160nm, a *p*-well peak doping and depth of about $8x10^{17}cm^{-3}$ and 1.8μm respectively, an *n*-well peak doping and depth of about $1x10^{17}cm^{-3}$ and 4μm respectively, source and drain peak doping and depth of about $1x10^{21}cm^{-3}$ and 0.15μm respectively, a *p*-type halo implant of Boron under the condition of 25 KeV, $5x10^{13}cm^{-3}$, a 15° tilt, and quad rotation, a channel width in the range of 0.15μm to 0.25μm, and a channel length in the range of 0.25μm to 0.35μm. It will be appreciated that these characteristics are illustrative, and may vary depending on the application and the fabrication process. Transistors designed for CHE programming need not be optimized for the generation of channel-initiated secondary electrons, as described in, for example, J.D. Bude et al., EEPROM/Flash Sub 5.0V Drain-Source Bias Hot Carrier Writing, IEDM Technical Digest, 1995, p. 989-991.

**[0042]** Memory array programming operations for CHE-programmed serial flash memory preferably are performed on successive sets of multiple bits, and preferably on a byte-by-byte basis with all the bits of a byte being programmed simultaneously, and multiple bytes or an entire pate of memory being programmed in successive bytes. Memory array erase operations preferably are performed on multiple pages, one or more selected sectors, or the entire memory array. The selected memory cells for these memory

operations are accessed by placing appropriate voltages on the word, bit and source lines of the selected memory cells, as well as on the diffusion wells in which the selected memory cells reside. The non-selected cells have different combinations of voltages, including in some cases word, bit and source lines that are brought to ground potential or left floating, which prevent the operation from occurring on them.

**[0043]** Consider, for example, the NOR array 100. All of the transistors 111-114 are erased simultaneously by placing an erase select voltage on the word line $R_1$, an erase support voltage on all of the column lines $C_1$, $C_2$, $C_{(N-1)}$ and $C_N$, and an erase support voltage on the *p*-well in which the channel is formed. Any suitable technique may be used for erase convergence. If for purposes of illustration one does not wish to erase the transistors 121-124, 131-134 and 141-144, the word lines are brought to ground potential. As another example, the transistors 111-114 are selected for programming by placing a program select voltage on the word line $R_1$ and grounding the source line $S_1$. The column lines $C_1$, $C_2$, $C_{(N-1)}$ and $C_N$, carry a program support voltage or ground potential, depending on the data to be programmed into the memory array 100. If for purposes of illustration one does not wish to program the transistors 121-124, 131-134 and 141-144, the word lines $R_2$, ..., $R_{(M-1)}$ and $R_M$ would normally be brought to ground potential. It will be appreciated that the voltage levels depend on not only the type of memory cell, but also on the specific characteristics of the stacked gate memory cell. For example, a reduction in the oxide thickness between the floating gate and the channel permits reductions in the source-drain voltage.

**[0044]** As another example, consider the virtual ground array 200. Access to individual memory cells on a selected word line for reading and programming is obtained by applying appropriate voltages to the bit lines to avoid disturbing the threshold voltage of the cell adjacent to and on the same word line with the cell being read or programmed, in a manner well known in the art. For example, memory cell 232 is accessed for reading or programming by bringing the word line 230 high, the bit line 211 ($BL_{+1}$) high (illustratively 1.5 volts for reading, and 4.5 volts for programming), and the bit line 212 ($BL_0$) low (illustratively 0 volts). Disturbance of the threshold voltage of the adjacent

cells 231 and 233 is avoided by bringing the bit line 210 ($BL_{+2}$) high and the bit line 213 ($BL_0$) low.

**[0045]** We have found that flash memory that uses CHE for programming multiple bytes and entire pages suffers disturbance of the floating gate potential during programming operations in the memory cells controlled by the unselected word lines and sharing the same bit line or bit lines as the cell or cells being programmed. Memory cells in the memory arrays 100 and 200 are impressed with four different sets of voltages during programming, two sets being impressed on the selected cells depending on the whether data is being programmed therein, and the other two sets being impressed on unselected cells depending on their location in the memory array. These various sets of voltages are illustrated in FIG. 5, FIG. 6, FIG. 7 and FIG. 8 for the NOR memory array 100, but are applicable in principle to other types of arrays such as the virtual ground array 200.

**[0046]** FIG. 5 shows illustrative voltages on a selected memory transistor that is being programmed: the gate voltage $V_G = 10.5V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 4.5V$ and the *p*-well substrate voltage $V_P = 0V$. Majority carriers flow in the channel, as indicated by the arrow with the black arrowhead. Additionally, channel hot electrons are generated and injected into the floating gate, as indicated by the arrow with the white arrowhead. The resulting threshold voltage is illustratively $V_T = 5.5V$. This is a satisfactory condition.

**[0047]** FIG. 6 shows illustrative voltages on a selected memory transistor that is not being programmed: the gate voltage $V_G = 10.5V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 0V$ and the *p*-well substrate voltage $V_P = 0V$. A similar condition exists in the virtual ground array 300, except that the source and drain of the memory cell adjacent the memory cell being programmed are pulled up. No channel current flows, and no channel hot electrons are generated. This is a satisfactory condition.

**[0048]** FIG. 7 shows illustrative voltages on a non-selected memory transistor that shares a bit line with a selected memory transistor that is being programmed: the gate voltage $V_G = 0V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 4.5V$ and the *p*-well substrate voltage $V_P = 0V$. Some majority carrier flow exists in the channel, as indicated

by the arrow with the black arrowhead. Channel hot electrons are generated, as indicated by the arrow with the white arrowhead. Since the control gate is at zero potential, it does not attract the hot electrons. However, some of the hot electrons will nonetheless become injected into the floating gate after a prolonged time, thereby disturbing the erased state of this transistor. Although the effect is small for each programming operation, the cumulative effect for a very large and very dense memory array can be sufficiently large to raise the transistor's threshold voltage $V_T$ so that it reads like a programmed cell. This is an unsatisfactory condition, and is found in the memory arrays 100 and 200.

**[0049]** FIG. 8 shows illustrative voltages on a non-selected memory transistor that shares a bit line with a selected memory transistor that is not being programmed: the gate voltage $V_G = 0V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 0V$ and the *p*-well substrate voltage $V_P = 0V$. No channel current flows, and no channel hot electrons or secondary hot electrons are generated. This is a satisfactory condition.

**[0050]** The unsatisfactory condition illustrated in FIG. 7 is improved by placing a negative bias on the gates of at least the unselected transistors having differential voltages on their sources and drains. An illustrative value is about minus 1.5 volts, although the precise voltage depends on the type of memory transistor and its specific characteristics. The various sets of illustrative voltages that result are shown in FIG. 9 and FIG. 10.

**[0051]** FIG. 9 shows illustrative improved voltages on a non-selected memory transistor that shares a bit line with a selected memory transistor that is being programmed: the gate bias voltage $V_G = -1.5V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 4.5V$ and the *p*-well substrate voltage $V_P = 0V$. Some majority carrier flow exists in the channel, as indicated by the arrow with the black arrowhead. Channel hot electrons are generated, as indicated by the arrow with the white arrowhead. However, since the control gate is at a negative bias, it repels the hot electrons and relatively few become injected into the floating gate. The erased state of this transistor is not significantly disturbed, even if the transistor remains unselected for a great number of programming operations common for a very large and very dense memory array. This is a satisfactory condition.

**[0052]** FIG. 10 shows illustrative improved voltages on a non-selected memory transistor that shares a bit line with a selected memory transistor that is not being programmed: the gate bias voltage $V_G = -1.5V$, the source voltage $V_S = 0V$, the drain voltage $V_D = 0V$ and the *p*-well substrate voltage $V_P = 0V$. No channel current flows, and no channel hot electrons are generated. The negative bias on the gate has no adverse effect. This is a satisfactory condition.

**[0053]** Exemplary program characteristics are shown in FIGS. 15 and 16. FIG. 15 is a graph of threshold voltage vs. time that shows program speed with different values of $V_G$ (points corresponding to $V_G = 9.5V$ are marked by the symbol ◊, points corresponding to $V_G = 10V$ are marked by the symbol □, and points corresponding to $V_G = 10.5V$ are marked by the symbol Δ), and indicates that cells can be programmed to $V_T \geq 5.5V$ in less than about 1 μsec with $V_G = 10.5V$, $V_D = 4.5V$, and $V_P = 0V$. FIG. 16 is a graph of drain current vs. drain voltage, and shows program current with different gate voltages (points corresponding to $V_G = 9.5V$ are marked by the symbol ◊, and points corresponding to $V_G = 10.5V$ are marked by the symbol o), and fixed $V_D = 4.5V$, $V_S = 0V$, and $V_P = 0V$. With $V_G = 10.5V$, a program current of about 320 μA is achievable. For flash memory with 8 bits per byte, the simultaneous programming of all of the bits of a byte is feasible, requiring about 2.5 mA of programming current. The one page programming time for a low density part of about 16 MB having a page size of 1024 bits is about 128 μsec, while the one page programming time for a high density part of about 32 MB having a page size of 2048 bits is about 256 μsec. Depending on the fabrication process, even less programming current per cell may be achievable, permitting either a lower total programming current or the simultaneous programming of even more bits.

**[0054]** It will be appreciated that the time required for programming a page of a serial flash memory using CHE programming is improved over the time required for programming a page of a serial flash memory using FN programming, which typically is on the order of about 5 ms and could be 7 ms and greater. The amount of some supporting circuitry is also reduced. A serial flash memory using FN programming requires a page of bit latches to store an entire page of data while the page of memory is

being programmed. In contrast, flash memory using CHE programming requires only 8 bit latches to store the data while each successive byte in the page is being programmed.

**[0055]** We have also found that the relatively thin tunnel oxide commonly found in the stacked gate type of flash memory transistor, including the type that is programmed using channel hot electron injection as well as the type that is programmed using channel-initiated secondary electron injection as disclosed in, for example, United Stated Patent No. 5,659,504, which issued August 19, 1997 (Bude et al., "Method and Apparatus for Hot Carrier Injection") and is incorporated herein by reference in its entirety, suffers a shortened lifetime because of the high fields that occur across the tunnel oxide during the FN erase operations. Memory cells in memory arrays such as the arrays 100 and 200 are impressed with three different sets of voltages during erase operations, one set being impressed on all of the selected cells and the other two sets being impressed on unselected cells depending on their location in the memory array. These various sets of voltages are illustrated in FIG. 11, FIG. 12 and FIG. 13 for a cell of the NOR memory array 100, but are applicable in principle to other types of arrays such as the virtual ground array 200. It also will be appreciated that the voltage levels depend on not only the type of memory cell, but also specific characteristics of the stacked gate memory cell and the application.

**[0056]** FIG. 11 shows illustrative voltages on a selected memory transistor that is being erased: the gate voltage $V_G = -12V$, the drain voltage $V_D = 6V$, the *p*-well substrate voltage $V_P = 6V$, and the *n*-well substrate voltage is 6V. The source voltage $V_S$ is left floating. Alternatively, the source voltage $V_S$ may be set to 6V and the drain voltage $V_D$ left floating, or both the source voltage $V_S$ and the drain voltage $V_D$ may be set to 6V. Electrons move by FN tunneling from the floating gate through the oxide to the channel and drain regions, thereby decreasing the $V_T$ of the transistor. The electron tunneling, which is generally parallel to the direction of the electrical field, is indicated by the arrows in FIG. 11. This is a satisfactory condition.

**[0057]** FIG. 12 shows illustrative voltages on a non-selected memory transistor: the gate voltage $V_G = 0V$, the drain voltage $V_D = 6V$, the *p*-well substrate voltage $V_P = 6V$, and the *n*-well substrate voltage is 6V. The voltage difference between the

channel/drain and the gate is insufficient to support FN tunneling. When the transistor is in an unprogrammed state, which is to say that the floating gate typically contains relatively few electrons, the electric field across the tunnel oxide is modest and does relatively little harm to the tunnel oxide, even if the memory undergoes many erase operations but the transistor itself is not erased. However, when the transistor is in a programmed state, which is to say that the floating gate contains many electrons and is highly charged, the electric field across the tunnel oxide is quite high. For example, for the stacked gate transistor of FIG. 1 holding a full charge, the voltage on the floating gate may be as greatly negative as minus 3 volts. The cumulative effect for a very large and very dense memory array can be sufficient, leading to degradation of the tunnel oxide. This is an unsatisfactory condition, which we refer to as erase disturb.

**[0058]** The unsatisfactory condition when the transistor is in a programmed or high $V_T$ state is improved by placing a positive bias on the gates of the unselected transistors in the memory array during erase operations. An illustrative value is about 2.5 volts to about 3 volts, that is $V_{CC}$, although the precise voltage depends on the type of memory transistor and its specific characteristics. The various set of voltages that results is illustrated in FIG. 13.

**[0059]** FIG. 13 shows illustrative voltages on a non-selected memory transistor having a positive gate bias: the gate bias voltage $V_G = 2.5V$, the drain voltage $V_D = 6V$, the *p*-well substrate voltage $V_P = 6V$, and the *n*-well substrate voltage is 6V. The voltage difference between the channel/drain and the gate remains insufficient to support FN tunneling. When the transistor is in an unprogrammed state, the electric field across the tunnel oxide is even smaller than in the FIG. 12 arrangement. When the transistor is in a programmed state, the electric field across the tunnel oxide is reduced by 2.5 volts, thereby bring it down to a lower field strength so that the cumulative effect for a very large and very dense memory array is acceptable. This is a satisfactory condition.

**[0060]** Exemplary erase characteristics are shown in FIG. 17, which is a graph of threshold voltage vs. time, for a stacked gate cell having a halo implant. Specifically, FIG. 17 shows erase speed for different values of $V_G$ (points corresponding to $V_G = -9V$ are marked by the symbol ◊, points corresponding to $V_G = -10V$ are marked by the

symbol o, and points corresponding to $V_G = -11V$ are marked by the symbol Δ), with $V_D = 6V$ and $V_P = 6V$ and with the *n*-well substrate voltage at 6V. With $V_G = -11V$, cells are erased to $V_T \leq 2V$ within about 300 μsec to about 500 μsec. FIG. 18 is a graph of threshold voltage vs. time, and shows the amount of erase disturb for four combinations: $V_G = 2.5V$ and $V_D = V_S = V_P = 6V$ (corresponding points marked by the symbol ◊); $V_G = 2.5V$ and $V_D = V_S = V_P = 7V$ (corresponding points marked by the symbol □); $V_G = 0V$ and $V_D = V_S = V_P = 6V$ (corresponding points marked by the symbol Δ); and $V_G = 0V$ and $V_D = V_S = V_P = 7V$ (corresponding points marked by the symbol X). The threshold voltage is seen to be essentially unchanged when using $V_G = 2.5V$ instead of $V_G = 0V$ during erase operations. FIGS. 17 and 18 are illustrative both of cells designed for CISEI programming as well as cells designed for CHE programming.

**[0061]** The counter biasing techniques described herein is not limited to the specific memory array architecture of FIGS. 3 and 4, but may be used in any array architecture, including virtual ground flash memory, NAND, NOR, and so forth, in which the individual unselected memory cells are subject to either voltage disturb during programming or to high electric fields across the tunnel oxide during erase.

**[0062]** The stacked gate transistor is read using any suitable technique and any suitable set of voltages, including page mode and single and multiple byte reading. Illustratively, the stacked gate transistor of FIG. 1 is designed to have a low $V_T$ of about 2.0V and a high $V_T$ of about 5.5V, although the precise voltage depends on the type of memory transistor and its specific characteristics.

**[0063]** To avoid any tendency for any memory cells in memory arrays such as the arrays 100 and 200 to experience read disturb due to hot electrons, which is similar to program disturb but on a smaller scale, a negative bias is placed on the gates of the transistors in the memory array that are not selected for reading. An illustrative value is about minus 1.5 volts, although the precise voltage depends on the type of memory transistor and its specific characteristics.

**[0064]** The illustrative set of voltages described above are summarized in the table shown in FIG. 14, which also includes illustrative voltages suitable for the memory array 200. The substrate *n*-well voltage $V_{NW}$ is also shown. It will be appreciated that the

particular voltages and voltage ranges set forth are illustrative and that satisfactory voltages different than the voltages and voltage ranges set forth in the table may be used. These voltages and variations thereof are generated and applied to the memory array using voltage multipliers and voltage dividers in a manner well know in the art.

**[0065]** The description of the invention and its applications as set forth herein is illustrative and is not intended to limit the scope of the invention. Variations and modifications of the embodiments disclosed herein are possible, and practical alternatives to and equivalents of the various elements of the embodiments are known to those of ordinary skill in the art. These and other variations and modifications of the embodiments disclosed herein may be made without departing from the scope and spirit of the invention.